coreservlets.com – Hadoop Course

Oozie

In this exercise, you will have a chance to implement Oozie workflow and get acquainted with deploying, executing and monitoring Oozie MapReduce workflows.

Approx. Time: 60 minutes

Perform

1. Start Oozie
2. Implement, deploy, and execute Oozie workflow. This workflow is composed of two steps where step 2 uses the output of step 1. The first step of this workflow calculates the number of occurrences for each distinct token, and the second step computes the average number of occurrences for each start letter. Your output should look something like this:

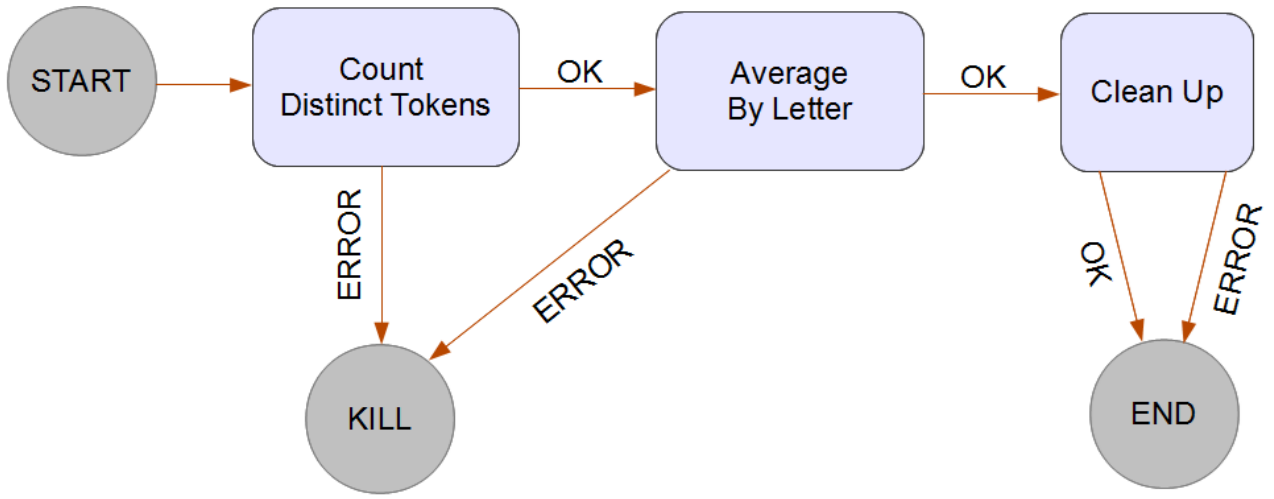
```
$ hdfs dfs -cat /training/playArea/oozieExerciseWorkflow/part-r-00000
"     1
#     1
...
...
?     2
A     8
B     5
C     3
...
...
y     17
z     1
```

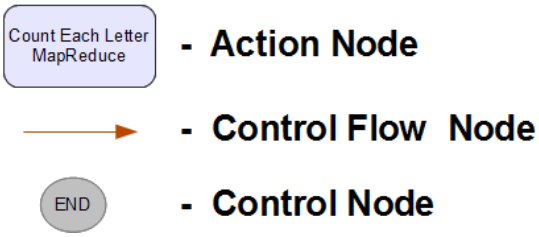
First MapReduce Step:

- mapper: `oozie.workflows.CountDistinctTokensMapper`
- combiner: `oozie.workflows.CountDistinctTokensReducer`
- reducer: `oozie.workflows.CountDistinctTokensReducer`

Second MapReduce Step:

- mapper: `oozie.workflows.AverageByLetterMapper`
- combiner: `oozie.workflows.AverageByLetterReducer`
- reducer: `oozie.workflows.AverageByLetterReducer`





You can find a template in the Exercises project but will need to implement

`workflow.xml` and `job.properties`

under

`/src/main/resources/oozie/workflows/`

Run `mvn package` for Exercises project to assemble and copy job's contents to

$`PLAY_AREA/oozie/oozie-exercise-workflow`

you can then use those artifacts to deploy and run this Oozie workflow. You will need to use Oozie's Process Definition Language to configure intermediary directory that the first job will use as output and second job as input. Make sure the workflow deletes output and intermediary directories prior starting as well as deleting intermediary directory upon success.

Solution

1. To start Oozie exeucute start script:

```
$ cd $OOZIE_HOME/bin
$ ./oozie-start.sh
```

2. You will find `workflow.xml` and `job.properties` in the Solutions project:

```
src/main/resources/oozie/workflows/workflow.xml
src/main/resources/oozie/workflows/job.properties
```

   To deploy and execute, run `mvn package` for Exercise project then copy Oozie application directory to HDFS and utilize `oozie` command to initiate the workflow:

```
$ cd $PLAY_AREA/oozie
$ hdfs dfs -rm -r oozie-exercise-workflow
$ hdfs dfs -put oozie-exercise-workflow
$ oozie job -config oozie-exercise-workflow/job.properties -run
```

   As an example of how to script provisioning and starting of a workflow take a look at:

```
$TRAINING_HOME/scripts/runOozieWorkflow.sh
```